Appendices

Many details are omitted in the main text because of space concerns; we present relevant details here.

1. Appendix A: Reproducibility statement
2. Appendix B: Ethics statement
3. Appendix C: Details of training data aggregation
4. Appendix D: Training details and hyperparameters
5. Appendix E: Standard deviations for few-shot results
6. Appendix F: Example zero-shot predictions on our evaluation tasks.
7. Appendix G: Additional text-type results
8. Appendix H: Generalized zero-shot learning setting

A. Reproducibility Statement

We ensure reproducibility of our results by releasing our datasets (TreeOfLife-10M and Rare Species), data pre-processing code, training code, evaluation code, code to generate all figures (Figs. 2 and 3), and pre-trained model weights. With these resources, anyone with sufficient compute resources can download the original data, then reproduce the pre-processing, training, and evaluation. For those with limited compute, the pre-trained model weights enable full reproducibility of our evaluation results (§4).

We provide DOIs as permanent references to our new digital assets:

- TreeOfLife-10M: doi:10.57967/hf/1972
- Rare Species: doi:10.57967/hf/1981
- BioCLIP: doi:10.57967/hf/1511
- Code: doi:10.5281/zenodo.10895871

B. Ethics Statement

We are not aware of any major ethical issues that arise from our work. BioCLIP is further pre-trained from the original CLIP model; many of the same concerns (class design, surveillance, etc.) apply; however, these concerns are discussed in great detail in Radford et al. [69], so we will focus on addressing these concerns as they relate to the biological addition provided in BioCLIP.

How could BioCLIP affect endangered species–does BioCLIP or TreeOfLife-10M pose a threat by aiding poachers? Though BioCLIP leads to improved automatic species classification, it does not include specific geographic information such as GPS coordinates. Furthermore, animal conservation status is not included during training.

Could BioCLIP have a negative impact on biologists? BioCLIP is designed to combine visual cues with an established taxonomic hierarchy to aid in scientific discovery. Concerns regarding over-reliance on model predictions is a warning that accompanies many–if not all–contemporary models and is not unique to BioCLIP. The goal is for BioCLIP to aid biologists in their work, not to replace them. As such, it is important for users to retain that understanding/context when applying BioCLIP to downstream tasks.

C. Training Data Aggregation

We aggregate images and labels from the iNat21 training data, Bioscan-1M's, and data downloaded from EOL. While every image has at least one taxonomic rank labeled, full taxonomic hierarchies and common names are scraped on a best-effort basis from metadata providers, including iNaturalist (iNaturalist Taxonomy DarwinCore Archive), Encyclopedia of Life (eol.org) and Integrated Taxonomic Information System (ITIS) (itis.gov).

We create a lookup between scientific name and taxonomic hierarchy and a lookup between scientific name and common name. We populate these lookups using the following sources in order of descending prioritization, as earlier sources are considered more authoritative. That is, if a duplicate appears in a later source, it is superseded by the higher priority source: Bioscan-1M metadata, EOL aggregate datasets: information retrieved using EOL page IDs with the pages API, which checks for a match in the ITIS hierarchy for higher-level taxa standardization (setting aside homonyms for proper linkage). The full list of taxa and vernacular names provided by iNaturalist and the iNat21 training set class names were maintained. From here, any taxa that could not be resolved using these sources were fed through the Global Names Resolver (GNR) API. Overall we were able to achieve 84% full taxa labeling for images in TreeOfLife-10M, for context, 10% of TreeOfLife-10M is only labeled down to the family rank (Bioscan-1M), thus, genus-species information is not available.

Despite our efforts, we discovered after training that some hemihomonyms were mislabeled at higher-level taxa (family up to kingdom). This impacts approximately $0.1 - 0.2\%$ of our data. We are in the process of developing a more robust solution to taxonomic labeling which will also account for re-naming (as is currently in process for many bird species). We intend to release a patch alongside the solution.

D. Hyperparameters & Training Details

Tabs. D1 and D2 contain our training hyperparameters for the different models. Tab. D2 notes the different epochs at which we had the lowest validation loss, as evaluated using the CLIP objective on the validation split of TreeOfLife-10M (even for the TreeOfLife-1M models). We will release our training code upon acceptance.

We trained a hierarchical classification model in §4.4. Python pseudocode for the training objective is in Listing 1. We will publicly release full training code upon acceptance.

Hyperparameter	Value
Architecture	ViT-B/16
Max learning rate	1×10^{-4}
Warm-up steps	1,000
Weight Decay	0.2
Input Res.	224×224

Table D1. Common hyperparameters among all models we train.

Dataset	Text Type	Batch Size	Epoch
TreeOfLife-10M	Mixture	32K	100
iNat21 Only	Mixture	16K	65
TreeOfLife-1M	Common	16K	86
	Scientific		87
	Taxonomy		87
	Sci+Com		87
	Tax+Com		86
	Mixture		91

Table D2. Hyperparameters that differ between the various models we train. We use a smaller batch size and only 1M examples for our text-type ablation because of limited compute.

```python
import torch.nn.functional as F

def forward(vit, heads, images, h_labels):
  """
  vit: vision transformer.
  heads: linear layers, one for each taxonomic
         rank.
  images: batch of input images
  h_labels: hierarchical labels; each image has
            7 labels
  """
  img_feats = vit(images)
  h_logits = [head(img_feats) for head in heads]
  losses = [F.cross_entropy(logits, label)
    for logits, labels in zip(h_logits, h_labels)]
  return sum(losses)
```

Listing 1. Python code to calculate the hierarchical multitask objective. Each image has 7 class labels: one for each taxonomic rank. The ViT calculates dense features for each image, then each taxonomic rank has its own linear layer that produces logits. By summing the losses, the ViT learns to produce image features that are useful for classifying images at multiple taxonomic ranks.

E. Standard Deviation of Main Results

Tabs. E3 and E4 show the accuracy with standard deviation over five runs on the test sets presented in Tab. 2. Since we randomly select the training examples from the datasets for few-shot, accuracies vary based on which examples are train examples and which are test examples. However, the variation is small enough that our conclusions in §4.5 still hold. Zero-shot results are deterministic and have no variation.

F. Example Predictions

Figs. F1 and F2 show BioCLIP and CLIP zero-shot predictions on all ten evaluation tasks. We randomly pick examples from each dataset that BioCLIP correctly labels and example that CLIP incorrect labels but BioCLIP correctly labels. BioCLIP performs well on a variety of tasks, including out-of-distribution images (Plankton, Medicinal Leaf) and mixes of scientific and common names (PlantVillage, PlantDoc).

G. More Results of Text-Type

We investigated the effects of text-type during training and testing in §4.3 using the Rare Species task. We present zero-shot results for all text-types on all tasks using the same procedure as in §4.2, where we use whatever taxonomic+common if available, otherwise whatever text-type is available.

H. Generalized Zero-Shot Learning

Chao et al. [17] introduced *generalized zero-shot learning*, where a model must label images of unseen classes from a set of both seen and unseen labels. We pick out a set of 400 seen species from TreeOfLife-10M using the same methodology as we used for the Rare Species task. We classify the same images from the Rare Species task using this set of 800 labels (a mix of seen and unseen). CLIP and OpenCLIP achieve 23.0% and 18.2% top-1 accuracy, while BioCLIP achieves 26.0% top-1 accuracy in this challenging GZSL setting.

Model	Birds 525	Plankton	Insects	Insects 2	Rare Species
One-Shot Classification					
CLIP	43.7 ± 0.26	25.1 ± 0.71	21.6 ± 1.05	13.7 ± 1.09	28.5 ± 0.65
OpenCLIP	53.7 ± 0.52	32.3 ± 0.63	23.2 ± 1.58	14.3 ± 0.67	29.2 ± 0.64
Supervised-IN21K	60.2 ± 1.02	22.9 ± 0.84	14.7 ± 1.38	14.4 ± 0.90	28.0 ± 0.77
DINO	40.5 ± 0.96	**37.0 ± 1.39**	23.5 ± 1.49	16.4 ± 0.78	31.0 ± 0.89
BioCLIP	71.8 ± 0.47	30.6 ± 0.77	**57.4 ± 2.4**	**20.4 ± 1.28**	**44.9 ± 0.73**
– iNat21 Only	**74.8 ± 0.89**	29.6 ± 0.82	53.9 ± 0.97	19.7 ± 0.80	36.9 ± 1.02
Five-Shot Classification					
CLIP	73.5 ± 0.37	41.2 ± 1.01	39.9 ± 0.86	24.6 ± 0.90	46.0 ± 0.33
OpenCLIP	81.9 ± 0.25	52.5 ± 0.83	42.6 ± 0.82	25.0 ± 0.83	47.4 ± 0.34
Supervised-IN21K	83.9 ± 0.15	39.2 ± 1.66	32.0 ± 1.90	25.4 ± 2.13	47.3 ± 0.41
DINO	70.9 ± 0.34	**56.9 ± 1.61**	46.3 ± 1.37	28.6 ± 1.59	50.1 ± 0.47
BioCLIP	90.0 ± 0.12	49.3 ± 1.14	**77.8 ± 0.81**	**33.6 ± 0.74**	**65.7 ± 0.43**
– iNat21 Only	**90.1 ± 0.08**	48.2 ± 1.24	73.7 ± 0.65	32.1 ± 1.97	55.6 ± 0.16

Table E3. Accuracy with standard deviation of five runs on animals and rare species in Tab. 4

Model	PlantNet	Fungi	PlantVillage	Med. Leaf	PlantDoc
One-Shot Classification					
CLIP	42.1 ± 3.40	17.2 ± 0.78	49.7 ± 2.53	70.1 ± 2.83	24.8 ± 1.61
OpenCLIP	45.1 ± 3.40	18.4 ± 1.26	53.6 ± 0.79	71.2 ± 3.58	26.8 ± 1.45
Supervised-IN21K	46.7 ± 6.30	16.9 ± 2.32	**62.3 ± 2.28**	58.6 ± 4.45	27.7 ± 2.86
DINO	30.7 ± 3.79	20.0 ± 1.53	60.0 ± 2.15	79.2 ± 2.74	23.7 ± 2.48
BioCLIP	64.5 ± 2.15	**40.3 ± 3.00**	58.8 ± 2.83	**84.3 ± 1.90**	**30.7 ± 1.75**
– iNat21 Only	**67.4 ± 4.54**	35.5 ± 2.93	55.2 ± 1.58	75.1 ± 1.16	27.8 ± 1.31
Five-Shot Classification					
CLIP	65.2 ± 1.25	27.9 ± 2.54	71.8 ± 1.46	89.7 ± 1.45	35.2 ± 1.59
OpenCLIP	68.0 ± 0.86	30.6 ± 1.26	77.8 ± 1.28	91.3 ± 0.85	42.0 ± 1.32
Supervised-IN21K	70.9 ± 2.45	30.9 ± 2.64	**82.4 ± 1.53**	82.3 ± 3.81	44.7 ± 2.26
DINO	50.3 ± 3.20	34.1 ± 2.87	82.1 ± 1.31	94.9 ± 1.30	40.3 ± 2.32
BioCLIP	**85.6 ± 1.79**	**62.3 ± 1.82**	80.9 ± 1.04	**95.9 ± 1.07**	**47.5 ± 1.35**
– iNat21 Only	84.7 ± 1.24	55.6 ± 2.61	77.2 ± 0.68	93.5 ± 1.13	41.0 ± 1.75

Table E4. Accuracy with standard deviation of five runs on plants and fungi in Tab. 4

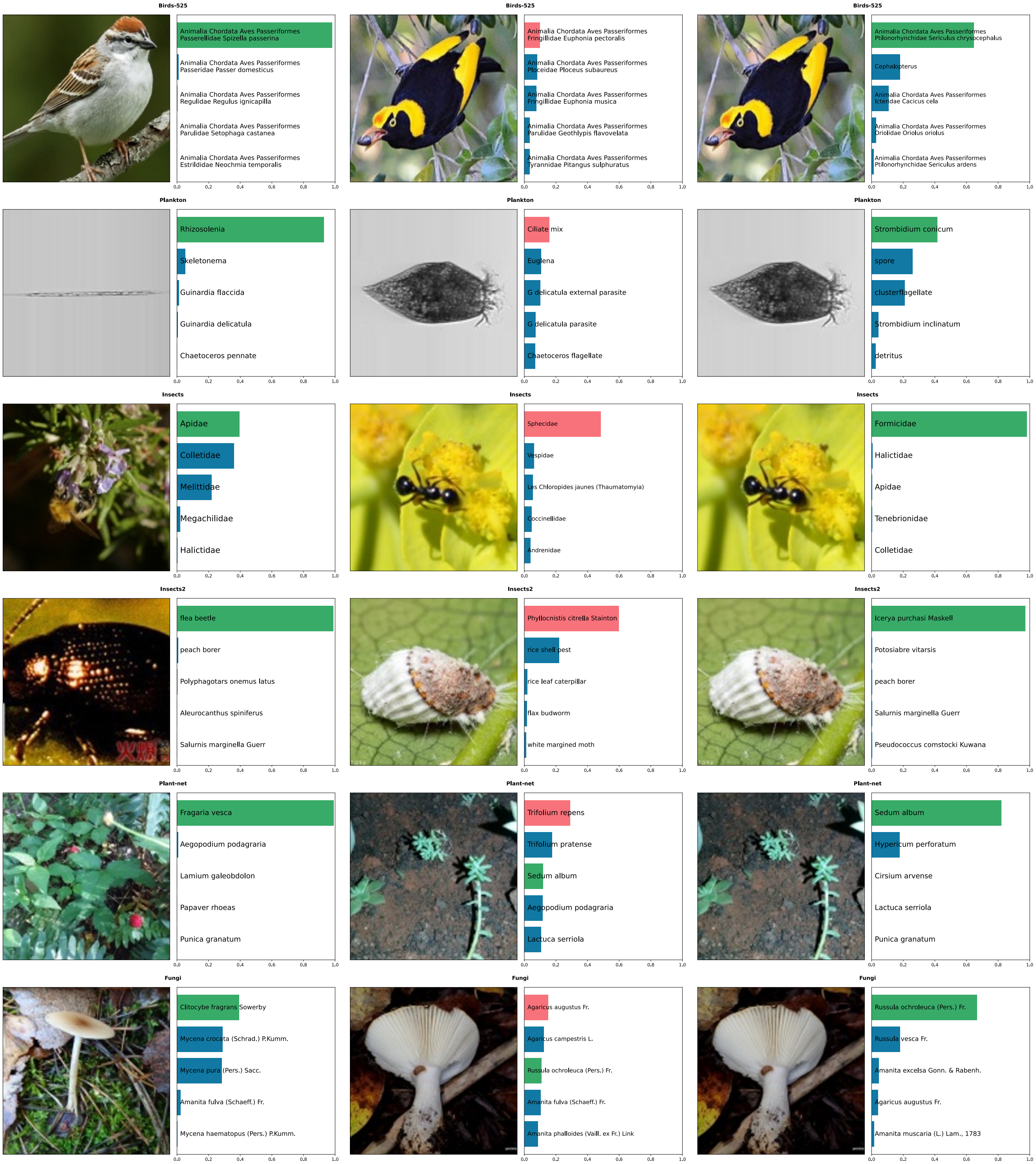

Figure F1. Example predictions for BIOCLIP and CLIP on Birds 525, Plankton, Insects, Insects2, PlantNet and Fungi tasks. Ground truth labels are green; incorrect predictions are red. Left: Correct BIOCLIP predictions. Center, Right: Images that CLIP incorrectly labels, but BIOCLIP correctly labels.

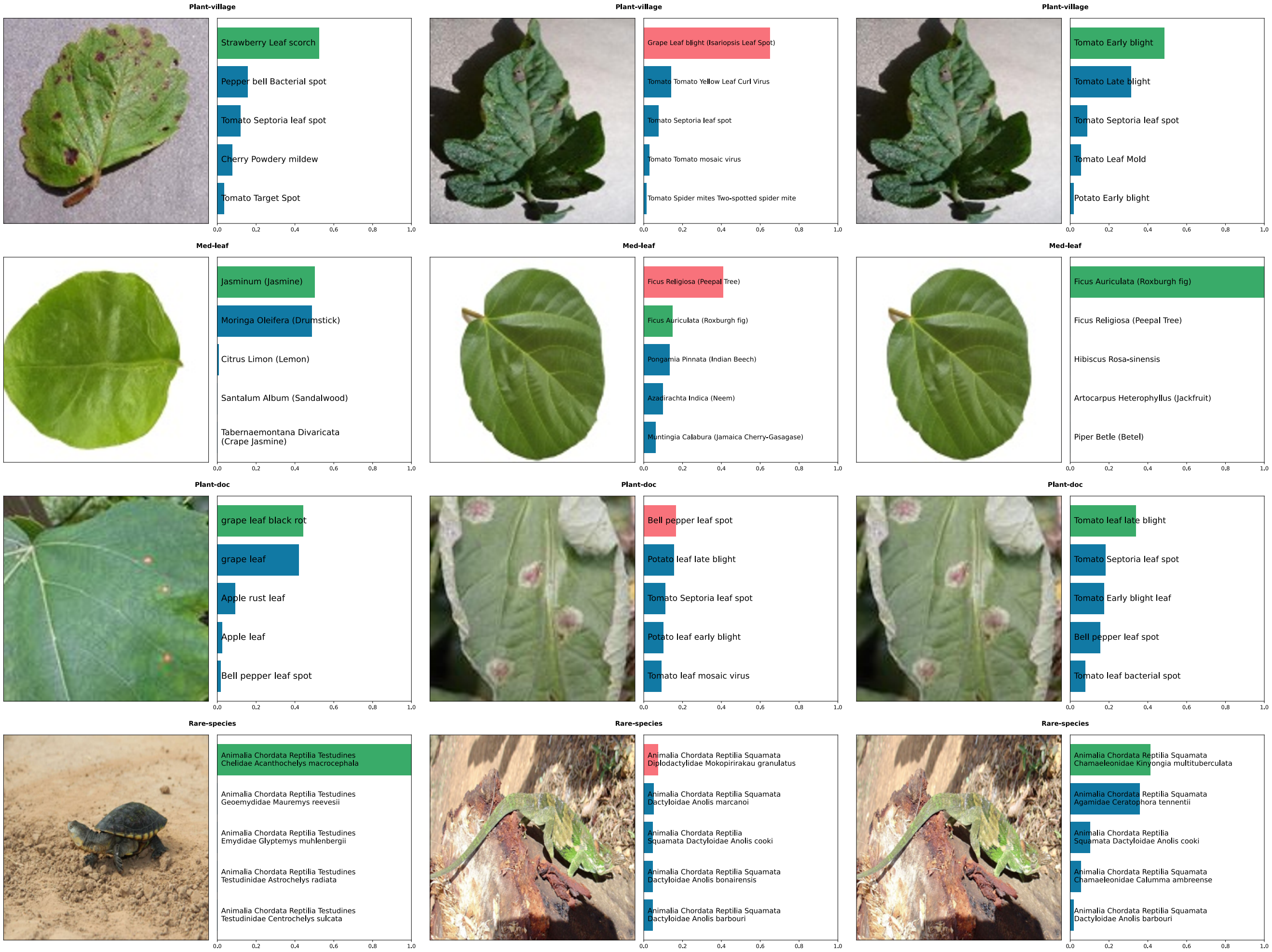

Figure F2. Example predictions for BIOCLIP and CLIP on PlantVillage, Medicinal Leaf, PlantDoc and RARE SPECIES. Ground truth labels are green; incorrect predictions are red. Left: Correct BIOCLIP predictions. Center, Right: Images that CLIP incorrectly labels, but BIOCLIP correctly labels.

	Animals				**Plants & Fungi**							
Training Text Type	Birds 525	Plankton	Insects	Insects 2	PlantNet	Fungi	PlantVillage	Med. Leaf	PlantDoc	Rare Species	Mean (Δ)	
Random Guessing	0.2	1.2	1.0	1.0	4.0	4.0	2.6	4.0	3.7	0.3	2.2	
Common	58.5	**4.4**	15.8	13.3	45.2	20.7	10.7	15.4	19.6	24.9	22.8	−10.1
Scientific	59.7	3.8	18.7	11.0	84.8	**35.3**	12.5	20.3	13.9	22.3	28.2	−4.7
Taxonomic	62.7	2.2	25.1	8.7	70.4	29.0	8.8	18.4	12.8	26.6	26.4	−6.5
Sci+Com	60.2	2.2	19.2	12.6	71.5	24.8	17.6	**21.5**	20.0	28.0	27.7	−5.2
Tax+Com	60.2	2.0	27.4	11.6	68.4	19.2	10.4	19.5	15.8	30.4	26.4	−6.5
Mixture	**65.1**	3.5	**30.6**	**17.3**	**86.3**	32.8	**19.9**	18.7	**24.5**	**30.9**	**32.9**	–

Table G5. Zero–shot classification top-1 accuracy for different text-types used during training. **Bold** indicates best accuracy. All models use the same architecture (ViT-B/16 vision encoders, 77-token text encoder) and are trained on the same dataset (TREEOFLIFE-1M). Δ denotes the difference in mean accuracy with "Mixture", which is the text-type we used for BIOCLIP.